OntoGene at CALBC II and Some Thoughts on the Need of Document-Wide Harmonization

Clematide, S ; Rinaldi, Fabio ; Schneider, G

OntoGene at CALBC II and Some Thoughts on the Need of Document-Wide Harmonization

Simon Clematide, Fabio Rinaldi, Gerold Schneider

Institute of Computational Linguistics, University of Zurich
{siclemat,rinaldi,gschneid}@cl.uzh.ch

Introduction

The OntoGene group has developed several syntax-based approaches for relation mining in the molecular biology domain, especially for the detection of mentions of protein-protein interactions. The effectiveness of these approaches has been validated by participation to shared evaluations, such as BioCreative II.5 [1] and III [2], or BioNLP event extraction task [3]. For the first CALBC challenge [4], the dictionary-based term recognizer originally developed for BioCreative II.5 has been adapted to the needs of large scale annotation. This system makes use of an efficient dictionary-based longest-match lookup procedure for the annotation of token sequences, which includes a flexible normalization to deal with surface variants of the terms stored in the dictionaries. The text tokens undergo the same normalizations as the original dictionary terms, thus allowing direct comparison of the normalized version of a textual candidate term with the normalized version of a reference term.

For the second CALBC challenge, we retained this dictionary-based engine for candidate term generation. However, as our results for protein and gene recognition of the first CALBC challenge showed, the bias towards high recall and lower precision, which works well for protein-protein interaction detection, needs remedy. Therefore, we filtered the candidate terms of our dictionary-based engine by a statistical hidden Markov Model (HMM). In particular, we trained a "First-Best Named Entity Chunking Model" using the LingPipe framework [5] for each of the 4 basic semantic types in the training corpora. For this supervised learning step we worked with the training corpora of CALBC I and II, as well as with the GENETAG corpus [6]. In order to benefit from the combination of a dictionary-based longest-match recognizer and a statistical chunker, we filtered the candidate terms for the final submission by the following rule: discard all candidate terms where the HMM chunker does not predict the begin of a named entity. We did not require exact correspondence of the end of terms deliberately, because we observed a slight bias towards shorter terms in the case of the HMM chunker.

Technical Details and Evaluations for Our Participation in Task A of CALBC II

First, we split the huge corpora into smaller slices in order to speed up the processing by massive parallelization. Then, we tokenized using Lingpipe's biomedical HMM tokenizer. The HMM chunker was applied separately for each semantic type we had trained it for. From this step, only a "begin-of-term" marker was retained for each term chunk, which ensures that the dictionary-based term recognizer excludes matches containing them. The term recognizer used the following external dictionaries in addition to the terms we extracted from the CALBC training corpora (I and II):

- UniProt for proteins and genes (prge): 826,901 terms (incl. CALBC terms)
- PharmGKB for diseases (diso): 36,080 terms (incl. CALBC terms)
- PharmGKB for chemical substances and drugs (ched): 43,997 terms1 (incl. CALBC terms)
- NCBI Taxonomy (and own resources) for species (spe): 903,880 terms (incl. CALBC terms)

Table 1 contains an overview of the filtering effect of the HMM chunker on the output of the dictionary-based term recognizer. The mismatch between both methods is rather high for all categories, except for the recognition of species. However, in a 10-fold cross-validation experiment with the HMM chunker on the CALBC II training corpus using an exact boundary recognition criterion, recall, precision, and F1-measure were not as high as one would expect (see Table 2). One reason may be that a simple HMM model generally performs worse than what can be expected from techniques as Conditional Random Fields [7].

1 The ChEBI 3 star level database was not used for the submissions because of an unfortunate configuration error of our system.

Another reason may be that the term annotation in the harmonized corpus is not as consistent as it should be. We tried to roughly quantify this effect by using the following one-sense-per-abstract hypothesis [8]: for each abstract in the CALBC II training corpus, all token sequences of annotated terms were collected. Then, in the same (!) abstract we searched for occurrences of term token sequences that were not annotated as terms of the same type. Under the one-sense-per-abstract hypothesis each unannotated occurrence counts as a false negative. Table 3 gives an overview of how many missing terms we can expect for each semantic type. This corresponds roughly to the results in Table 2.

Type	DTR	HMM Chunker	Validated	in %
prge	869,550	535,959	493,572	57%
ched	741,059	472,097	405,104	55%
diso	644,717	416,654	347,711	54%
spe	494,183	393,040	365,470	74%

Table 1: Amount of validated terms of dictionary-based term recognizer (DTR) by HMM NER chunker on 175k article set

Type	Recall	Precision	F1-Measure
spe	88%	82%	85%
diso	79%	73%	76%
prge	75%	57%	65%
ched	72%	50%	59%

Table 2: 10-fold cross-validation experiment on CALBC II training corpus using only the HMM NER chunk

Even if the one-sense-one-abstract hypothesis does not hold strictly (e.g. “in” should not be considered a protein in the whole abstract if it appears once as one), there is still room for improvement of document-wide harmonization of annotations (e.g. in one abstract “immunoglobulin” appears 6 times as a term, yet it is missed another 4 times).

Table 4 contains our official results evaluated against the consensus annotation from the CALBC project partners. These results show that simple HMM filtering in combination with dictionary-based term recognition improves in particular in the case of the more difficult problems, namely the recognition of protein/genes and chemical substances.

Type	Total Terms	Missing	in %	Miss. Types
prge	327,090	84,727	26%	6621
ched	165,098	36,604	22%	2355
diso	230,459	17,143	7%	987
spe	247,529	13,940	6%	718

Table 3: Total terms in CALBC II training corpus and missing terms according to one-sense-one-abstract hypothesis

Type	R (rank)	P (rank)	F (rank)	S
spe	88% (2)	86% (2)	87% (2)	14
diso	74% (5)	80% (5)	77% (4)	15
prge	83% (1)	67% (11)	74% (2)	18
ched	77% (2)	69% (7)	72% (1)	14

Table 4: Official results for recall (R), precision (R) and F1-measure (F) with the corresponding rank in parens. Column S contains the total number of submissions for each semantic type.

Funding: The OntoGene group is supported by the Swiss National Science Foundation (grant 105315-130558/1) and by NITAS/TMS, Text Mining Services, Novartis Pharma AG, Basel, Switzerland.

References

[1] F. Rinaldi, G. Schneider, K. Kaljurand, S. Clematide, T. Vachon, and M. Romacker, “OntoGene in BioCreative II.5,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 7, 2010, pp. 472-480.

[2] F. Rinaldi, G. Schneider, S. Clematide, M. Romacker, and T. Vachon, “OntoGene (Team 65): preliminary analysis of participation in BioCreative III,” *BioCreative III workshop*, 2010.

[3] K. Kaljurand, G. Schneider, and F. Rinaldi, “UZurich in the BioNLP 2009 Shared Task,” *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, Boulder, Colorado: Association for Computational Linguistics, 2009, pp. 28-36.

[4] S. Clematide, F. Rinaldi, and G. Schneider, “OntoGene in CALBC,” *First CALBC Workshop*, 2010, pp. 30-31.

[5] B. Carpenter, “LingPipe for 99.99 % Recall of Gene Mentions,” *Proceedings of the Second BioCreative Challenge*, 2007, pp. 2-4.

[6] L. Tanabe, N. Xie, L.H. Thom, W. Matten, and W.J. Wilbur, “GENETAG: a tagged corpus for gene/protein named entity recognition,” *BMC Bioinformatics*, vol. 6, 2005, p. S3.

[7] K. Hara, “Towards automatic biomedical entity annotation by reducing error propagation,” *First CALBC Workshop*, 2010, pp. 35-37.

[8] W.A. Gale, K.W. Church, and D. Yarowsky, “One sense per discourse,” *Proceedings of the workshop on Speech and Natural Language*, Association for Computational Linguistics, 1992, pp. 233-237.